Around the world in three alternations

Szmrecsanyi, Benedikt; Grafmiller, Jason; Heller, Benedikt; Röthlisberger, Melanie

DOI:
10.1075/eww.37.2.01szm

Document Version
Peer reviewed version

Citation for published version (Harvard):
Szmrecsanyi, B, Grafmiller, J, Heller, B & Röthlisberger, M 2016, 'Around the world in three alternations: Modeling syntactic variation in global varieties of English', *English World-Wide*, vol. 37, no. 2, pp. 109-137. https://doi.org/10.1075/eww.37.2.01szm

Around the world in three alternations:
modeling syntactic variation in varieties of English

Short title: Around the world in three alternations

Benedikt Szmrecsanyi (KU Leuven)

Jason Grafmiller (KU Leuven)

Benedikt Heller (KU Leuven)

Melanie Röthlisberger (KU Leuven)

Abstract

We sketch a project that marries probabilistic grammar research to scholarship on World Englishes, thus synthesizing two previously rather disjoint lines of research into one unifying project with a coherent focus. This synthesis is hoped to advance usage-based theoretical linguistics by adopting a large-scale comparative and sociolinguistically responsible perspective on grammatical variation. To highlight the descriptive and theoretical benefits of the approach, we present case studies of three syntactic alternations (the particle placement, genitive, and dative alternations) in four varieties of English (British, Canadian, Indian, and Singapore), as represented in the International Corpus of English. We report that the varieties studied share a core probabilistic grammar which is, however, subject to indigenization at various degrees of subtlety, depending on the abstractness of the syntactic patterns studied.

Keywords: probabilistic grammar, probabilistic indigenization, syntax, variation, dative alternation, genitive alternation, particle placement, multivariate analysis

Acknowledgments

Generous financial support from the Research Foundation Flanders (FWO, grant # G.0C59.13N) is gratefully acknowledged. We thank Gerold Schneider and Hans-Martin Lehmann from the University of Zurich for access to the Dependency Bank 2.0. Constructive comments by two anonymous referees and the editors have made this a better paper. Thanks go to Christy Thanh Linh Ha for a close reading of the manuscript. The usual disclaimers apply.

1. Introduction

This is a programmatic paper reporting on an on-going research project entitled "Exploring probabilistic grammar(s) in varieties of English around the world", which is situated at the crossroads of research on English as a World Language, usage-based theoretical linguistics, variationist linguistics, and cognitive sociolinguistics. It specifically marries the spirit of the PROBABILISTIC GRAMMAR FRAMEWORK (e.g., Bresnan 2007), which assumes that grammatical knowledge is experience-based and partially probabilistic, to research along the lines of what we call here the ENGLISH WORLD-WIDE PARADIGM (e.g. Schneider 2007; Mesthrie and Bhatt 2008), which is concerned with the sociolinguistics of post-colonial English-speaking communities around the world. The overarching objective is to understand the plasticity of the probabilistic knowledge of English grammar, on the part of language users with diverse regional and cultural backgrounds. Tapping into a large corpus of World Englishes and utilizing primarily quantitative analysis and modeling techniques, the aim is to probe the probabilistic factors that constrain syntactic variation in varieties of English. In this paper, we will address the following research questions:

1. Do the varieties of English we study here share a core probabilistic grammar?
2. Can ecology account for probabilistic similarity between varieties of English – for example, do we find a split between native and non-native varieties of English?
3. Do the alternations under study differ in terms of their probabilistic sensitivity to variety effects?

The project is innovative in that we synthesize two previously rather disjoint lines of research into one unifying project. In doing so, we hope to inject new methodological and theoretical ideas into research on varieties of English, and to provide the Probabilistic Grammar framework with a challenging empirical testing ground.

To showcase the descriptive and theoretical benefits of the approach, we present case studies that explore three grammatical alternations – the particle placement alternation (1), the genitive alternation (2), and the dative alternation (3) – in four varieties of English (British, Canadian, Indian and Singapore English), which are covered in the International Corpus of English.

(1) a. verb-object-particle order (V-DO-P)
you can just [*cut*]$_{\text{verb}}$ [*the tops*]$_{\text{direct object}}$ [*off*]$_{\text{particle}}$ *and leave them.* [ICE-GB:S1A-007]

b. verb-particle-object order (V-P-DO)
[*Cut*]$_{\text{verb}}$ [*off*]$_{\text{particle}}$ [*the flowers*]$_{\text{direct object}}$ *as they fade*. [ICE-CAN:W2B-023]

(2) a. the *s*-genitive
[*Singapore*]$_{\text{possessor}}$'s [*small size*]$_{\text{possessum}}$ *meant it could be quick to respond to changes in economic conditions* [ICE-SIN:W2C-011]

b. the *of*-genitive
the [*size*]$_{\text{possessum}}$ *of* [*the eyes*]$_{\text{possessor}}$ *is to help them at night.* [ICE-GB:W2B-021]

(3) a. the ditransitive dative variant
That will give [*the panel*]$_{\text{recipient}}$ [*a chance*]$_{\text{theme}}$ *to expand on what they've been saying.* [ICE-GB:S1B-036]

b. the prepositional dative variant
[...] *and that gives* [*a chance*]$_{\text{theme}}$ [*to Bhupathy*]$_{\text{recipient}}$ *to equalise the points at thirty all.* [ICE-IND:S2A-019]

We find that the varieties we study do share a core probabilistic grammar which is, however, subject to indigenization at various degrees of subtlety, depending on the abstractness and the lexical embedding of the syntactic pattern involved. Second, the varieties we investigate tend to cluster along native/non-native (or ENL/ESL) lines, though we hasten to add that ENL/ESL patterns in a selection of only four varieties should not be over-interpreted.

2. Theoretical framework

Research on the scope and limits of variation within and across varieties of English around the world is booming. A shortcoming of research in this tradition, however, is an often primarily descriptive interest in the variable presence or absence of linguistic features, or a focus on "surfacy" usage frequencies of grammatical patterns and markers. But while feature inventories and usage frequencies are no doubt interesting, they do not address the most interesting part of the story: Does language users' grammatical knowledge differ across varieties of English, and if so, to what extent? A focus on linguistic knowledge would go beyond mere description and link research on varieties of English to recent advances in linguistic theory, such as usage- and experience-based models of language.

With this in mind, we propose to explore differences in the hidden probabilistic constraints that fuel syntactic variation within and across varieties of English. A 'hidden' probabilistic constraint is a constraint that is not tied to surface material but to a more or less subtle stochastic generalization about usage, which language users implicitly know about (a fact that can be shown in experiments -- see, e.g., Bresnan 2007), such as the principle of end-weight (place long constituents after short constituents; see e.g. Behagel 1909; Wasow and Arnold 2003). Usage-based methodologies are well-suited to uncover such knowledge by exploring contextual characteristics of linguistic features in large corpora sampling naturalistic text and speech.

In this context, we specifically apply the idea of a dynamic probabilistic grammar (Bybee and Hopper 2001; Bod, Hay, and Jannedy 2003; Gahl and Garnsey 2004; Gahl and Yu 2006) to the realm of cross-varietal variation in World Englishes. To model language users' probabilistic knowledge, we rely on the variation-centered, usage- and experience-based probabilistic grammar framework developed by Joan Bresnan and collaborators (e.g. Bresnan 2007; Bresnan et al. 2007; Bresnan and Ford 2010). The work we report in this paper thus builds on two key assumptions:

1. Grammatical variation is sensitive to multiple and sometimes conflicting probabilistic constraints, be they formal, semantic, or contextual in nature (Bresnan 2007, 75). Such constraints, like the principle of end-weight, influence linguistic choice-making in subtle ways which may remain invisible unless analyzed quantitatively.

2. Grammatical knowledge must have a probabilistic component, for the likelihood of finding a particular linguistic variant in a particular context in a corpus has been shown to correspond to the intuitions that speakers have about the acceptability of that particular variant, given the same context (Bresnan and Ford 2010).

These assumptions and our methodology on the whole are broadly compatible with work in modern variationist sociolinguistics (see, e.g., Labov 1982; Tagliamonte 2001).

3. Varieties, methods, and data

We tap into the International Corpus of English (ICE) (see http://ice-corpora.net/ice/), which samples a range of registers in a number of varieties of English from around the world. In this paper, we will be concerned with syntactic variation in the following four varieties of English:

British English (henceforth: BrE), as sampled in ICE-GB – *the* Inner Circle (Kachru 1992) variety. Note that ICE-GB samples fairly standard language, as opposed to regional dialect speech.

Canadian English (henceforth: CanE) , as sampled in ICE-CAN – another Inner Circle variety that is fairly similar to American English (which is not currently fully covered in ICE). CanE has reached phase 5 ("differentiation") in Schneider's (2007) Dynamic Model.

Indian English (henceforth: IndE), as sampled in ICE-IND – an Outer Circle variety which Mukherjee (2007) considers a "steady state" phase 4 ("endonormative stabilization") variety.

Singapore English (henceforth: SgE), as sampled in ICE-SIN – another Outer Circle variety that has reached phase 4.

Thus, the varieties selected for analysis in this paper represent a healthy mix of variety types (native/ENL/Inner Circle versus indigenized L2/ESL/Outer Circle) which are spoken in different world regions and represent different evolutionary stages.

To explore particle placement, dative, and genitive choices as exemplified in (1) – (3), we adopt the variationist methodology and restrict attention to "alternate ways of saying 'the same' thing" (Labov 1972, 188). Therefore, we will be interested in only those particle placement, genitive, and dative occurrences where the competing variant could have been used. We explain how such "interchangeable" variants were identified in the corpus material in the relevant sections below.

Once the interchangeable variants were identified, we annotated the datasets for a number of predictors. In this particular paper, we restrict attention to predictor variables that can be annotated automatically. For each alternation, measures of givenness, thematicity, type-token ratio, and frequency were automatically annotated using a Perl script, with only very minor manual correction. Givenness captures whether a noun had been mentioned recently in the discourse: a constituent was coded as 'given' if its head noun (lemma) was mentioned in the 100 words prior to the actual occurrence, and as 'new' otherwise. Thematicity – the extent to which a word represents the topic or "theme" of a text (Osselton 1988) – was calculated by determining the total number of occurrences of a head noun in the text in which it occurs. If a noun is related to the overall theme of a text it will tend to be mentioned relatively often, thus its thematicity will be relatively high. Type-token ratio

was calculated over the 100-word context surrounding each dative/genitive/particle variant, i.e. 50 words on either side (see Hinrichs and Szmrecsanyi 2007, 457). If an observation was located towards the end of a corpus file, our script considered more of its preceding context, and vice versa, to ensure a context of 100 words in total. For the overall frequency of a constituent head noun, we consulted the corpus of Global Web-based English (GloWbE) (Davies and Fuchs 2015): our script automatically extracted the relative frequencies of the head nouns in the subsections of GloWbE that match their varietal origin. For example, frequency of a noun from ICE-SIN was defined as the relative frequency of that noun within the Singaporean component of GloWbE. In addition, each observation was annotated for two external variables: VARIETY (BrE vs. CanE vs. IndE vs. SgE), and a 4-level coding of GENRE: 'spoken monologue' vs. 'spoken dialogue' vs. 'printed text' vs. 'non-printed text'.

Based on these general predictors as well as other, alternation-specific factors (discussed below), we explore the effects of internal and external variables on the syntactic outcomes using two different but complementary methods (see also Bernaisch, Gries, and Mukherjee 2014). First, we modeled the data using conditional inference trees, which predict outcomes by recursively partitioning the data into smaller and smaller subsets according to those predictors that co-vary most strongly with the outcome. Informally, binary splits in the data are made by trying to maximize the homogeneity or "purity" of the data partitions with respect to the values of the outcome (e.g. all *s*-genitives vs. all *of*-genitives). At each step, the dataset is recursively inspected to determine the variable (and its values) that makes the purest split in the data. This splitting process is repeated until no further split that significantly reduces the impurity of the data partitions can be found. The result is visualized as a flowchart-like decision tree. However, while conditional inference trees provide results that are relatively easy to interpret, they are nonetheless subject to a high degree of variability depending on the data from which they are generated, due to the fact that each split in the tree depends on the splits that precede it. Not only does this make it difficult to generalize to unseen data, but it is also the case that the trees sometimes fail to notice highly predictive effects. To obtain a more reliable measure of predictor importance, we also modeled the data using Conditional Random Forest (CRF) analysis, as implemented in the cforest() function in R's party package (Hothorn, Hornik, and Zeileis 2006; Strobl et al. 2007; Strobl et al. 2008). The CRF approach is based on conditional inference trees; however, it uses ensemble methods in a forest of trees built on randomly sampled subsets of the data to arrive at an aggregated estimate of a particular outcome's probability. Additionally, each tree is restricted to a random subset of predictors whose significance is assessed through random permutation tests. By amalgamating the results over the entire forest of trees, we obtain a model that is both highly accurate and robust to predictor multicollinearity and data overfitting – two persistent problems for more established techniques, e.g. logistic regression models. More important, the tree and forest method is able to explore complex interactions in ways that surpass regression models. In short, the CRF offers a reliable measure of the overall importance of each predictor, while the single tree method provides an elegant visualization of the complex interactions among predictors. Given that our main interest in this paper lies in the variable effects of different linguistic factors across varieties, i.e. interactions, these methods offer an advantage over regression models (which we are currently pursuing). For more discussion, see Tagliamonte and Baayen (2012) and Baayen et al. (2013).

4. Case studies

In this section, we present our case studies: the particle placement alternation (Section 4.1.), the genitive alternation (Section 4.2.), and the dative alternation (Section 4.3.).

4.1. Particle placement

4.1.1. Data selection and extraction

The particle verb dataset consisted of all interchangeable transitive particle verbs involving one of the following 10 particles (based on Gries 2003, 67–68): *around, away, back, down, in, off, out, over, on, up.* Tokens were extracted from the CLAWS7 tagged versions of the four ICE corpora using simple regular expression searches for strings within a 15-word span containing any verb followed by one of the particles listed above. After the initial extraction, the data were filtered to exclude tokens that did not involve genuinely interchangeable uses. Easily identified exclusion contexts included passive sentences, sentences with extracted direct objects, modified particles (*send the ball right back*), and names, titles, or other fixed phrases (e.g. *Take Me Out to the Ball Game*). Outside these contexts, special care was taken to distinguish between sentences involving preposition verbs (4), and those involving particle verbs (5) (see Quirk et al. 1985, Ch16; Biber et al. 1999, Ch5).

(4) *At Notre Dame the Archbishop **called on** the international community to oppose the law of blood* [ICE-GB:S2B-010]
(5) *He first **switched on** a large fluorescent tube to which the protection system had had been attached* [ICE-GB:S2A-041]

Despite their surface similarities with "true" particle verbs, preposition verbs are not interchangeable. We tested questionable cases against several diagnostic environments which are known to allow only preposition verbs (see Cappelle 2005, 78–81). These included intervening adverbs (*the state **called** repeatedly **on** them for assistance*), PP extraction (***On** whom did the state **call** for assistance?*), *it* clefts (*It was **on** the national guard that the state **called***), and repetition in coordination (*Did the state **call on** the local police or **on** the national guard?*). If the verb was acceptable in any of these contexts, it was considered a preposition verb and excluded. As a final heuristic, if the interchangeability of a specific token was still uncertain, we searched for uses of the alternate variant in either Google (restricted by country domain) or GloWbE.[1] If five or more observations of the alternate variant could be found, the token was accepted as interchangeable. This selection and filtering process resulted in a dataset of 5,414 interchangeable particle verb tokens (Table 1).

	CanE	BrE	IndE	SgE	Total
V-P-DO	718 (42.4%)	686 (45.9%)	696 (73.9%)	814 (63.4%)	2,914 (53.8%)
V-DO-P	975 (57.6%)	810 (54.1%)	246 (26.1%)	469 (36.6%)	2,500 (46.2%)
Total	1,693	1,496	942	1,283	5,414

Table 1: Distribution of transitive particle verb variants in four varieties of English

[1] We stress that in the present study, web searches were used only as a means of providing (some) independent evidence for a token's possible interchangeability in lieu of native speaker judgments (which are themselves not always reliable). No web data were included in the actual analysis.

4.1.2. Predictor variables

A substantial body of research has identified numerous features that influence the choice of particle verb variant, including, most notably, factors relating to the discourse accessibility and length – or 'heaviness' – of the direct object, as well as the semantic properties of the verb (see e.g. Gries 2003 and literature cited therein). In addition to the predictors mentioned in section 3, we included the following predictors specific to the particle placement alternation in our analysis:

- Pronominality of the direct object (DIROBJPRONOMINALITY) – whether or not the direct object was a personal or reflexive pronoun. Because pronominal direct objects occurred nearly categorically in the split order (V-DO-P) in our dataset (99.8%), we restricted the attention to only those tokens involving non-pronominal direct objects (*N* = 4,025).[2]
- Length in words of the direct object (DIROBJLENGTH). In our dataset, we found that tokens involving direct objects of six words or fewer constituted 90.3% of all observations. Of the 390 tokens with objects greater than six words in length, we found only three (0.8%) in the split order. We therefore further restricted the data to only those tokens with direct objects between one to six words long (*N* = 3,635).
- Definiteness of the direct object (DIROBJDEFINITENESS) – two-level coding for the definiteness ("definite" vs. "indefinite") following the criteria of Garretson et al. (2004).
- DIRECTIONAL PP. The presence of a directional PP following the target VP (*His lively tail was **whisking up** dust* [*from the ground behind him*] [ICE-IND:W2F-014]).

4.1.3. Results

The classification accuracy of the conditional inference tree, 73.8% (*N* = 3,635), is significantly better than the baseline accuracy of 69.4% we would obtain by simply always choosing the more common variant ($p_{binom} < 0.001$). As a more robust measure of classification accuracy, we also calculated the concordance statistic *C*, which represents the probability that the model will rank any randomly chosen observation of the joined variant (the "success" value in our binary model) higher than any randomly chosen observation of a split token.[3] The *C* statistic is an accuracy measurement independent of the baseline accuracy, and ranges from 0.5 (random chance) to 1 (perfect prediction), with values above 0.8 reflecting a model with relatively good explanatory power (F. E. Harrell 2001; Tagliamonte and Baayen 2012). For the present tree we obtained a *C* statistic of 0.66.

The resulting tree diagram for the particle verbs dataset is shown in Figure 1. Interpreting the tree is straightforward. Each node represents a split in the data into two subsets corresponding to the values shown on the connecting lines, based on the importance of those values in predicting the outcome. At each node the model chooses the factor (and values) that provides the most reliable predictions (at the customary significance level $\alpha = 0.05$) and splits the data accordingly. The topmost

[2] Impersonal pronouns (*everyone, something, no one,* etc.) and bare demonstratives were included in the dataset due to the fact that they often occur in the joined (V-P-DO) order, whereas personal pronouns do not.

[3] To calculate *C*, we used the somers2() function in the Hmisc package (F. E. J. Harrell 2014).

node thus represents the most predictive (binary) split in the data overall, with lower nodes representing the most reliable splits within the ensuing sub-regions of the data. The terminal nodes provide a barplot of observed proportions of the V-DO-P and V-P-DO variants, along with the total number of tokens observed in the corresponding subsets of the data.

The first thing to note is that not all predictors are present in Figure 1. Both DIROBJGIVENNESS and DIROBJTHEMATICITY are absent, suggesting that these factors play a minor role in predicting the choice of variant in the model. However, we note that while the tree in Figure 1 has been restricted to the uppermost four branching levels for the sake of visualization, when we allow unlimited splits we find numerous interactions of DIROBJTHEMATICITY (i.e. thematicity-based splits) farther down in the tree. The significant role of DIROBJTHEMATICITY is reflected more robustly in the results of the Conditional Random Forest analysis discussed below.

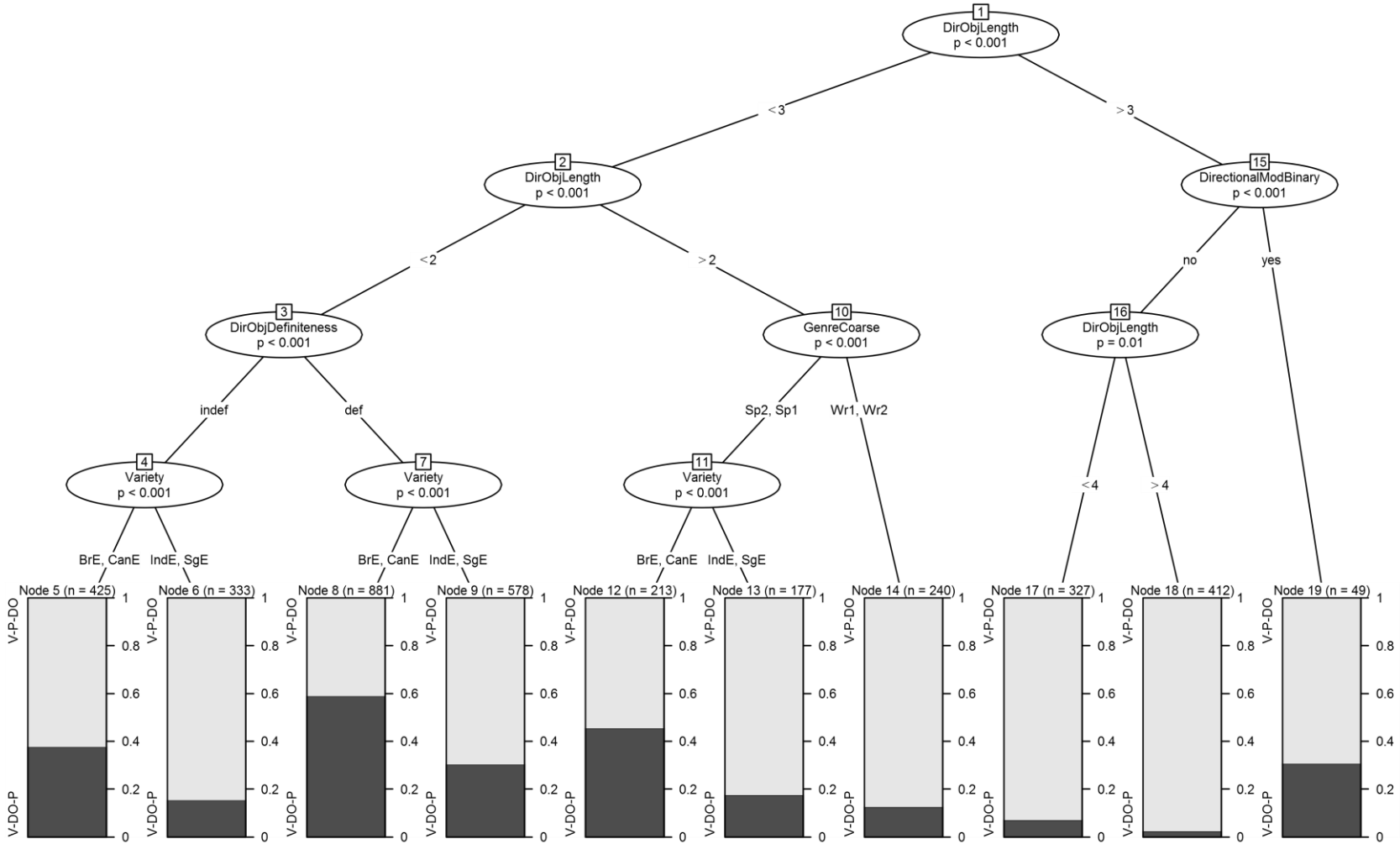

Figure 1: Conditional inference tree for particle placement. “Sp1” = spoken monologues, “Sp2” = spoken dialogues, “Wr1” = printed texts, “Wr2” = hand-written texts

Turning to the significant predictors, at Node 1, we see a major split into relatively short (three words or fewer) and relatively long (> three words) direct objects. Moving down from Node 1, the relative simplicity of the right side of the tree side suggests that there is comparatively little variability in particle placement among longer direct objects, at least with respect to the other predictors in the model. When the direct object exceeds three words in length, the split V-DO-P order is very unlikely, regardless of any other factors. The one notable exception is reflected in Node 15, which shows that even long(ish) direct objects are used relatively often (≈ 30%) in the split order when a directional PP is present. A further split in DIROBJLENGTH is also evident among the non-modified particle verb tokens (Node 16), though this effect is rather small, as indicated by the relatively minor change in the proportions of split tokens (the dark bars) between four-word direct objects and those of five to six words.

Proceeding down the left side, we find that this branch of the tree is more complex, with a further split at Node 2 between tokens with very short (one to two words) direct objects, and those with objects of moderate length (three words). Among the former subset, we see that definite NPs (Node 3) are significantly more likely to be used in the split order when the NP is very short. Among three-word direct objects however, the split order is only significantly more likely in spoken texts (Node 10): monologues ('Sp1') and dialogues ('Sp2'). DIROBJDEFINITENESS, however, is not a reliable predictor in these data.

Finally – and most pertinently given the topic of this paper – at the lowest branching level we find three splits by VARIETY (Nodes 4, 7, 11). Notably, each split separates the data in exactly the same way and in exactly the same direction: in all cases, the L1 varieties (BrE and CanE) show a significantly greater proportion of split variants than the L2 varieties (SgE and IndE). The location of these splits is important, as it shows that VARIETY is not as reliable a factor when the direct object is relatively long, which is reflected in the fact that there are no VARIETY-driven splits on the right side of the graph. Node 11 further reveals an interaction between GENRE and VARIETY among moderately long (three words) direct objects. Only in spoken English do we see an effect of VARIETY.

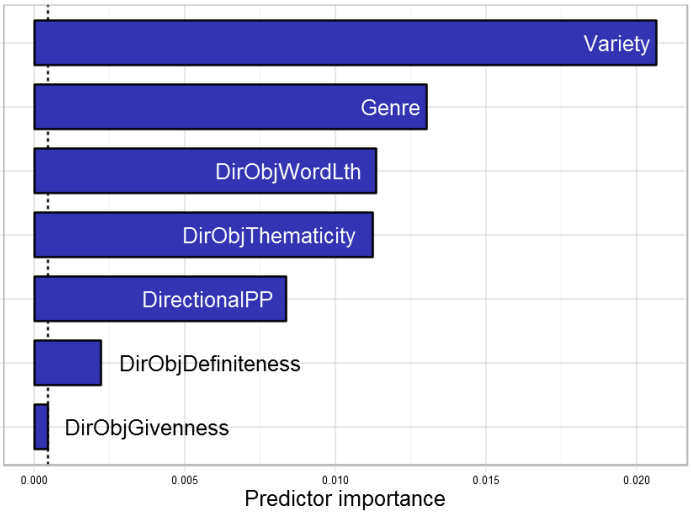

Figure 2: Predictor importance ranking for CRF analysis of particle placement

We now turn to a CRF analysis of the dataset. Since the procedure is, as the name implies, a random process, the analysis was run using two different random seeds to assess the stability of the results (Strobl et al. 2008). The CRF analysis performs significantly better than the single conditional inference tree, with a *C* statistic of 0.87 for both runs, and predictive accuracies of 80.7% and 80.5%, compared to the single conditional inference tree accuracy of 73.8% ($p_{binom} < .001$). In this instance we find that the CRF model is a substantial improvement over the single conditional inference tree model.

The explanatory power of individual predictors can be assessed by comparing the decrease in overall accuracy of the model when each predictor is removed. The greater the decrease, the more important the predictor. The relative ranking obtained from the CRF analysis is shown in Figure 2.[4] Here we see that more than any other predictor, VARIETY makes the largest contribution overall, followed by GENRE, DIROBJWORDLTH, DIROBJTHEMATICITY, and the presence of a directional PP. Neither DIROBJDEFINITENESS nor DIROBJGIVENNESS play much of a role however. This ranking is largely consistent with the results above, with a couple notable exceptions.

Observe, first, that it is somewhat surprising that the explanatory power of DIROBJWORDLTH would not rank higher than it does (3rd) in the CRF model given the preponderance of length-based splits observed in the inference tree in Figure 1 (see Nodes 1, 2, and 16). Second, DIROBJTHEMATICITY ranks somewhat higher than we expected in the CRF model, given that we found no splits at all in the single conditional inference tree model. We take these results as an illustration of (one of) the potential pitfalls of relying on a single conditional inference tree model, as discussed in section 3. The position of a predictor in a single tree is not always a reliable indicator of its relative importance overall, as a single tree only displays the largest binary split at any given point in the data. Because the CRF analyses and their variable importance rankings are based upon the conditional permutation of predictor variables over many, many trees, we can be reasonably confident that the rankings in Figure 2 reflect our best model of the relative explanatory power of each of our factors. The superiority of the CRF method is further reflected in its substantial improvement in classification accuracy (*C* = 0.87) over the single inference tree method (*C* = 0.66). Again, the usefulness of conditional inference trees lies in their ability to identify and illustrate the complex interactions among predictors, and less in their ability to rank predictors' overall importance.

Summarizing the results of our analysis of particle placement, we find that the internal factors we explored behave largely as expected. The split V-DO-P variant is favored when the direct object is short, definite, contextually salient, and when the VP is followed by a directional PP. Givenness, surprisingly, plays less of a role. Beyond this, several points deserve further comment.

[4] Importance rankings were obtained using the varimp() function in R's party package (Hothorn, Hornik, and Zeileis 2006; Strobl et al. 2008). The values on the x axis are not meaningful in their own right, but merely serve to allow comparisons between the importance of individual factors in the model.

First, more than any other internal factor, the length of the direct object significantly influences the choice of particle verb variant. Moreover, the effect of length is an incremental one: with each additional word in length, the split order becomes significantly less likely, to the extent that few other factors matter once the direct object exceeds three words in length. Second, results of the random forest analysis tell us that variety, genre (mode), and the length of the direct object each play an important role in predicting particle placement, while the conditional inference tree shows that there are important interactions among these factors.

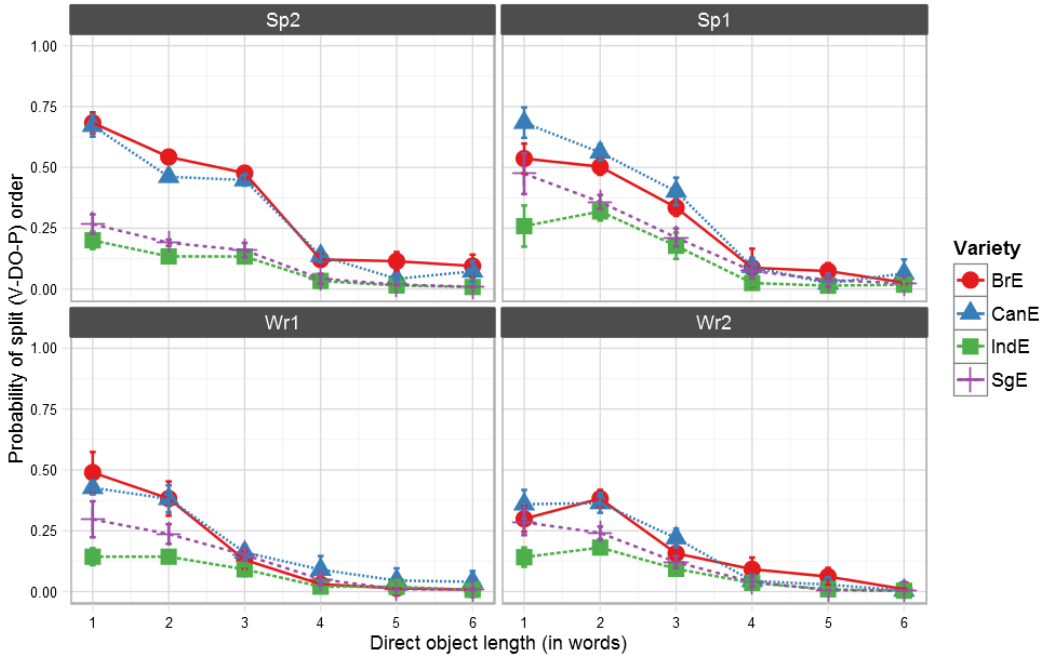

Figure 3: Predicted probabilities by direct object DIROBJLENGTH and VARIETY obtained from the Conditional Random Forest model (with 95% confidence intervals)

Cross-varietal differences only emerge when the direct object is relatively short (three words or fewer), and even then, such differences are found only in spoken data unless the direct object is very short (Figure 3).

(6) We would drive to the train station to ***pick*** dad ***up***. [CAN:S1A-042]
(7) It's difficult to imagine people ***picking*** women ***up*** and… [GB:S1A-020]
(8) somebody who are influenced by people who are speaking English ***picked up*** words. [IND:S1A-015]
(9) So I'm supposed to ***pick up*** somebody is it. [SIN:S1A-032]

This kind of interaction is entirely expected if weight-related effects are driven by language users' implicit desire to reduce processing demands (e.g. MacDonald 2013). We assume that speakers of all varieties are equally sensitive, on average, to the costs of accessing and processing longer constituents, and have roughly equivalent short-term memory capacities. Thus, it is not surprising to find what appears to be a consistent floor effect of end-weight across varieties: with direct objects beyond a threshold of three words, the split particle verb variant is very rarely used. We predict therefore that to the extent regional differences

in processing-driven factors exist, we should see them emerge only in those contexts where the processing load is relatively minimal. This is exactly what we find in our data. In such contexts, it is possible that variation in expectation-based processing effects (e.g. Levy 2008), which are grounded in usage probabilities, could explain the variation in end-weight effects exhibited across different varieties (see Bresnan and Ford 2010, 303–304). Such an explanation is entirely consistent with a usage-based probabilistic grammar framework.

4.2. The genitive alternation

4.2.1. Data selection and extraction

We were only interested in those genitive constructions that could, in principle, be expressed by both variants. For guidelines, see Rosenbach (2014) and literature cited therein.

Instances of interchangeable genitives were extracted in two stages: In stage one, we started with the extraction of all occurrences of the preposition *of*, and all occurrences of apostrophe (') + *s* and *s* + apostrophe. After that, we automatically filtered out all occurrences that either were not possessive constructions at all, or non-interchangeable genitive tokens, using a list specifying material that was not allowed to occur next to a candidate token. This list included, for example, *because* preceding *of* (i.e. *because of*), or *course* following *of* (i.e. *of course*). Part-of-speech-related constraints knocked out e.g. adjectives preceding or following *of*, which eliminated instances like *I'll be sick of being a god damn poor student* [ICE-NZ:S1A-051]. Finally, syntactic constraints discarded cases such as *I saw an eerie lightening of the sky* [ICE-NZ:W2B-011], in which the NP preceding *of* was not definite. After the filtering process, a total of 4,701 genitive tokens remained in the dataset (Table 2).

	CanE	BrE	IndE	SgE	Total
s-gen.	379 (34.5%)	285 (22.8%)	249 (18.8%)	302 (29.3%)	1,215 (25.9%)
of-gen.	721 (65.5%)	963 (77.2%)	1,074 (81.2%)	728 (70.7%)	3,486 (74.2%)
Total	1,100	1,248	1,323	1,030	4,701

Table 2: Distribution of genitive variants in four varieties of English

4.2.2. Predictor variables

As described in Section 3, givenness, thematicity (for the possessor: PORTHEMATICITY; for the possessum: PUMTHEMATICITY), type-token ratio (TYPETOKENRATIO), and frequency (PORHEADFREQ) were annotated automatically. Separate measures were computed for both the possessor and the possessum. For the three ratio-scaled factors (thematicity, type-token ratio, and frequency), we expected to find that comparatively high values would favor the *s*-genitive (Hinrichs and Szmrecsanyi 2007, Grafmiller 2014, *inter alia*). In addition to these factors, we included three genitive-specific predictors:

- Length of the possessor and possessum in orthographic words (PORWORDLTH and PUMWORDLTH). Previous research has shown that the *s*-genitive is favored when the possessor is short (Hinrichs and Szmrecsanyi 2007; Ehret, Wolk, and Szmrecsanyi 2014, inter alia). Compared to possessor length, the influence of possessum length on the alternation is less well understood, though the principle of end weight predicts that relatively short possessums should favor the *of*-genitive.
- Final sibilancy of the possessor (PORFINALSIBILANCY). The presence of a sibilant (/s/, /z/, /ʃ/, or /ʒ/) at the end of the possessor phrase, as in *The paradox's conclusion is that Achilles will never be able to catch up with the tortoise* [ICE-IND:W2B-021] makes the s-genitive less likely (Zwicky 1987).
- Noun phrase expression type of the possessor (PORNPEXPRTYPE). Noun phrase expression type was coded using part-of-speech information from the CLAWS-tagged versions of the ICE corpora. In the genitive alternation models, the levels of this factor are 'nc' (common noun), 'ng' (gerund), and 'np' (proper noun).

4.2.3. Results

The conditional inference tree in Figure 4 and the variable importance plot in Figure 5 show which of the factors under investigation are most important to explain the variability in the genitive data. The tree was limited to four branching levels. The predictive accuracy of the conditional inference tree amounts to 82.03% (baseline 74.15%), resulting in a *C* statistic of 0.72.

The first split in the conditional inference tree occurs between the levels of PORNPEXPRTYPE, separating genitives whose possessors are proper nouns ('np') from those with common noun ('nc') or gerund ('ng') possessors. Proper noun possessors are more likely to be realized as *s*-genitives than the other expression types, which can be observed by comparing the left side (nodes 5–12) of the tree to its right side (nodes 15–21). The overall importance of this factor, which exceeds the others by far (see Figure 5), can be explained by the fact that it overlaps to some extent with the predictive potential of possessor animacy since the proportion of animate referents is considerably higher in proper nouns than it is in common nouns (see the Conclusion section for a discussion of the animacy issue).

In both the left (nodes 5–12) and the right (nodes 15–21) branches of the tree, the second most prominent factor is the length of the constituents. Among common noun possessors, a subsequent split (node 2) divides instances involving shorter possessums of three words or fewer from those instances with longer possessums (> three words). Node 3 then concerns the values of possessor length, separating rather short possessors on the left (nodes 5–6) from longer ones on the right (nodes 8–9). Thus we find that when the possessor is a common noun, possessor length exerts its strongest influence only when the possessum is relatively short. Moreover, in such cases, if the possessor is longer than the possessum (node 9), genitives are almost exclusively realized as *of*-genitives. When *both* constituents are rather short however, the influence of a final sibilant enters the picture (node 5). For possessums longer than three words (right side branching from node 2), the interaction with possessor length does not matter as much; here, it is more important to know whether the

possessum is moderately long (three to five words, node 11), in which case both variants are almost equally likely, or very long (> five words, node 12), in which case the likelihood of an *s*-genitive realization rises to more than 80%.

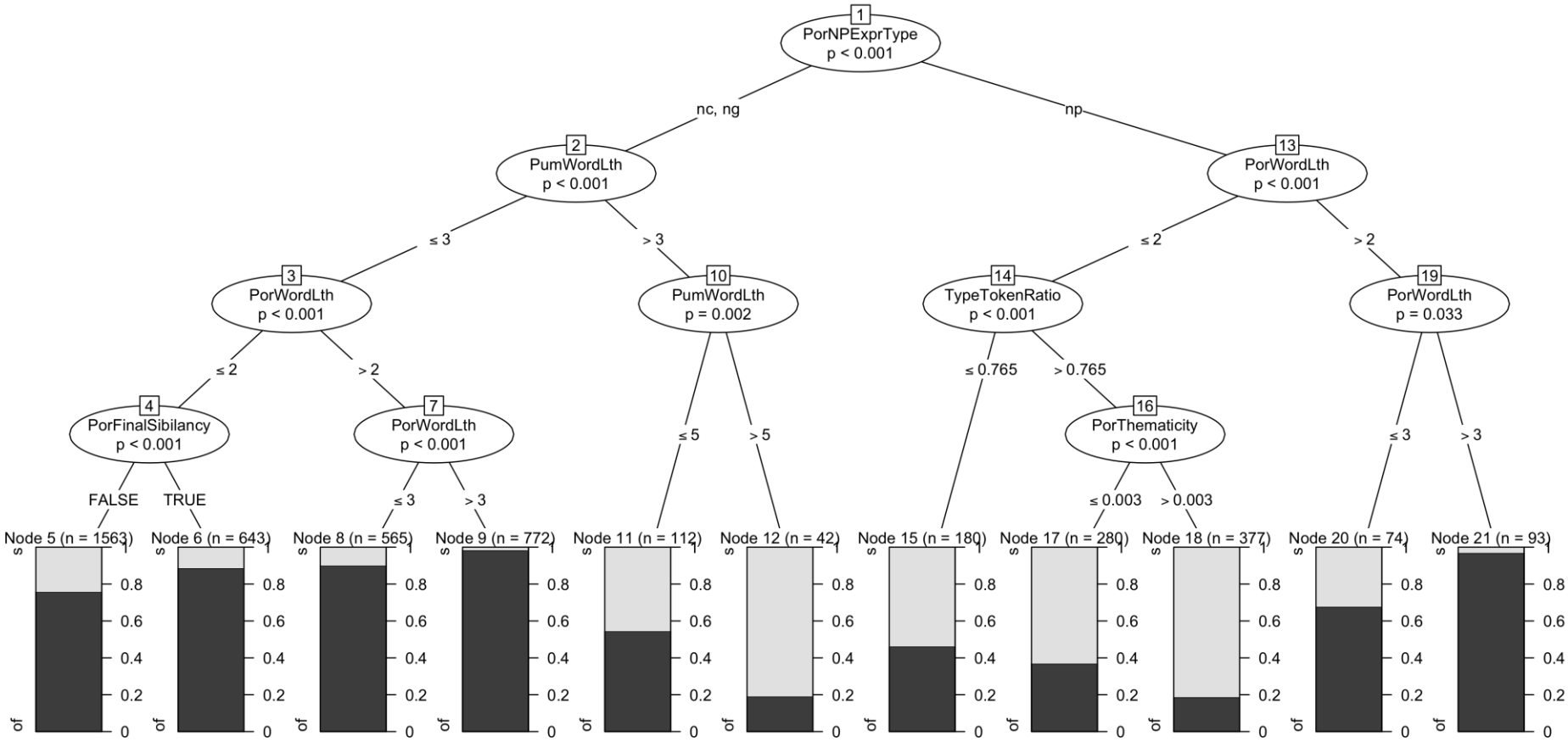

Figure 4: Conditional inference tree for genitive choice. “nc” = common noun, “ng” = gerund noun, “np” = proper noun

On the right branch of the tree, which contains variants with a proper noun possessor, the length of the possessor is more important than possessum length. Possessors longer than three words (node 21) are almost exclusively realized as *of*-genitives. If the possessor is between two and three words in length, *of*-genitives prevail by approximately 70% (node 20); if the possessor is short (one to two words), it is the *s*-variant that is more likely (nodes 15–18). Within this latter group, two other factors become important: TYPETOKENRATIO and PORTHEMATICITY. For both predictors, higher values – reflecting greater informational density and a higher degree of discourse salience for the possessor – further increase the likelihood of an *s*-genitive (nodes 17–18).

Although VARIETY does not appear in the top four levels of the conditional inference tree, if the tree is extended to five levels (not shown here), we do find a split along the lines of nativeness (i.e. CanE and BrE vs. IndE and SgE) for proper noun possessors with high values of TYPETOKENRATIO and PORTHEMATICITY (below node 18). Under such circumstances, the *s*-genitive is far more frequent in CanE and BrE (almost 90% of the cases) than in the other varieties (less than 60%). This suggests that the probability modification caused by PORTHEMATICITY (node 16) only affects the native varieties, but not IndE and SgE, which both use the *of*-genitive slightly more than 40% of the time, comparable to the level of node 15. Under node 5, we have a different split, which pits CanE and SgE against BrE and IndE. In this case, however, there is far less of a difference (< 10%) in genitive use than in the split below node 18 (around 30%).

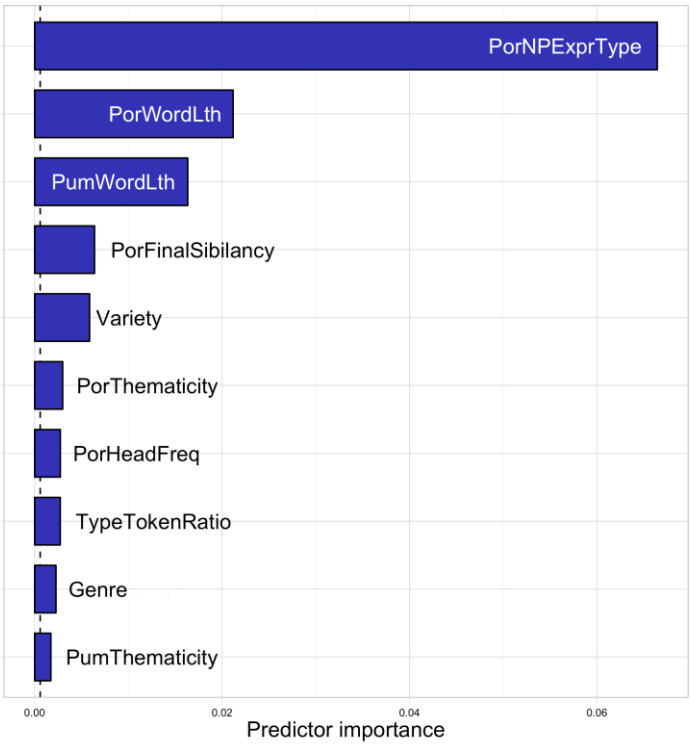

Figure 5: Predictor importance ranking for CRF analysis of genitive choice (displayed: 10 most important predictors)

For the genitive alternation, the CRF analysis paints a very similar picture to the conditional inference tree, but with a better statistical fit. The random forest model predicts 81% of all cases correctly and reaches a *C* statistic of 0.85, which is higher than the values of the conditional inference tree model. A second run yielded the same values. PORNPEXPRTYPE plays a dominant role, followed by the two length measures, which influence genitive choice along the lines of the principle of end weight. In fourth place in the variable importance plot we find PORFINALSIBILANCY, which also shows up in the tree. VARIETY comes in fifth, and thus

outperforms other predictor variable language-internal predictors such as thematicity, frequency, or type-token ratio, and the language-external predictor GENRE. However, type-token ratio and thematicity play a more important role in the right branch of the conditional inference tree than in the predictor importance plot from the random forest analysis, which points to possible interactions between variety and other predictors.

4.3. The dative alternation

4.3.1. Data selection and extraction

To extract dative occurrences, we compiled a list of dative verbs permitting both the ditransitive dative and the prepositional dative variant, based on data from a parsed version of the corpus material as well as the previous literature (see e.g. De Cuypere and Verbeke 2013; Schilk, Bernaisch, and Mukherjee 2012; Wolk et al. 2013). After incorporating this verb list into a perl script, we extracted all relevant dative tokens from the corpus, searching for strings of the form [*verb + NP/pronoun + ("to") + NP/pronoun*], the optional *to* capturing prepositional datives. Note that as is customary in the literature (e.g. Bresnan et al. 2007), we included both nominal and pronominal recipients and themes – unlike in the particle placement and genitive alternations, pronominal constituents in the dative alternation do not knock out particular variants. Also as in much previous dative alternation research, we then weeded out observations involving particle verbs, passivized verbs, elliptical structures, coordinated verbs, and clausal or non-overt constituents, as well as any cases that were not proper datives or genuinely variable. These included fixed expressions, constructions with spatial goals, and constructions that allowed a beneficiary reading. If it was unclear whether a particular dative token could be paraphrased by the other variant, we conducted a region-specific search in either Google or GloWbE to determine whether the paraphrase was attested. The dative dataset to be analyzed here spans 3,958 interchangeable datives (Table 3).

	CanE	BrE	IndE	SgE	Total
ditransitive dative	673 (72.6%)	642 (73.0%)	613 (56.3%)	772 (72.6%)	2,700 (68.2%)
prepositional dative	254 (27.4%)	237 (27.0%)	476 (43.7%)	291 (27.4%)	1,258 (31.8%)
Total	927	879	1,089	1,063	3,958

Table 3: Distribution of dative variants in four varieties of English

4.3.2. Predictor variables

In addition to the predictors listed in Section 3, we further include the following two predictors to model the dative alternation:

- Length in number of words of recipient (RECWORDLTH) and theme (THEMEWORDLTH). The dative alternation is a word order alternation, hence length effects play a crucial role in the ordering of the constituents (see e.g. Bresnan et al. 2007).
- NP expression type of recipient (RECNPEXPRTYPE) and theme (THEMENPEXPRTYPE). As with the genitive alternation, we annotated the recipient and theme for NP expression type, and

distinguished between six levels: common noun ('nc'), proper noun ('np'), personal pronoun ('pp'), impersonal pronoun ('iprn'), demonstrative ('dm'), and gerund ('ng').[5]

4.3.3. Results

Figure 6 shows the conditional inference tree for the dative alternation. Classification accuracy is 87.1%, which is significantly better compared to the baseline of 68.2% when always choosing the most frequent variant ($p_{binom} < 0.001$); the *C* statistic is 0.86. Again, the tree was limited to four branching levels.

First, note that the predictors related to givenness, thematicity, frequency, as well as GENRE and TTR are missing from the tree. That givenness and thematicity do not appear is probably due to the fact that these variables strongly correlate with other factors (e.g. length). Thematicity of the theme and recipient only appears at the bottom of the inference tree if we do not limit the tree to four branching levels.

Turning to the other variables, the most predictive factor in our inference tree is the length of the theme (Node 1). When THEMEWORDLTH is one word, we find several interactions between RECNPEXPRTYPE and THEMENPEXPRTYPE. From Node 8 we see that if recipient and theme are both expressed as personal pronouns, the prepositional variant is almost always used. If the recipient is not a personal pronoun and the theme is not an impersonal pronoun (Node 4), the prepositional variant is also preferred (e.g. *I give it to John*, but *I give John something*). Ditransitive variants are heavily favored when the recipient is realized as a personal pronoun and the theme is not (Node 7).

The right branch of the tree involves observations with themes longer than one word. Length measurements are the most important predictor variables: theme length and recipient length split the data down to the fourth level. Only then do RECNPEXPRTYPE and VARIETY come into play. If theme length is two to four words, and the recipient is one word long (Node 11), pronominal recipients favor the ditransitive variant (e.g. *I give him the book*) while the odds for a prepositional variant are higher if the recipient is not a personal pronoun (e.g. *I give the book to John*). When the theme is relatively long (> four words), the choice of variant again depends on the length and RECNPEXPRTYPE. Only when both recipient and theme are longer than four words (Node 21) do we find a proportionally greater amount of prepositional variants. We interpret this preference for prepositional datives in these contexts to be a reflection of Rohdenburg's (1996) 'Complexity Principle', i.e. the tendency for speakers to prefer the more explicit variant in cognitively more complex environments. When RECLENGTH is four words or fewer, the ditransitive variant is preferred – robustly so when the recipient is two words or fewer, and especially when it is a pronoun (Node 18). In the latter case, we nearly categorically find only ditransitive variants.

[5] Constituents that fell into none of these categories were labeled as 'O' for 'Other'.

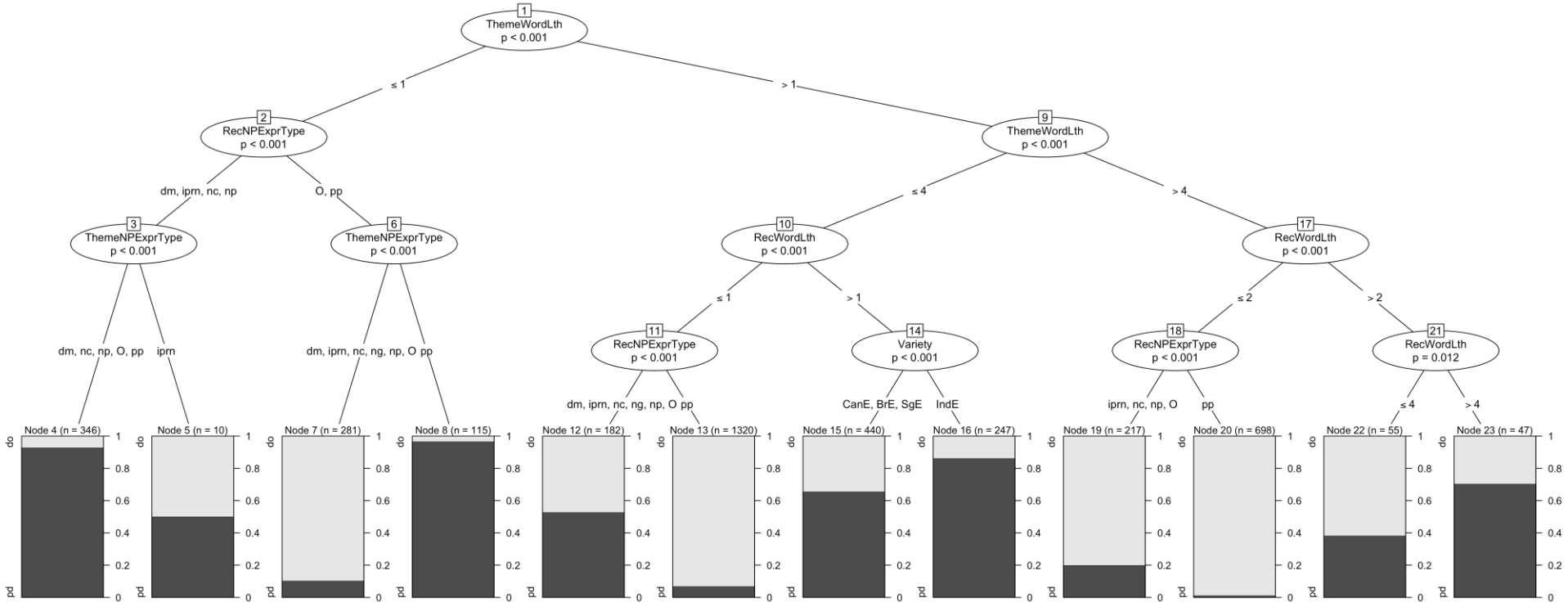

Figure 6: Conditional inference tree for the dative alternation. "nc" = common noun, "np" = proper noun, "pp" = personal pronoun, "iprn" = impersonal pronoun, "dm" = demonstrative, "ng" = gerund, and "O" = other

If we do not restrict the branching levels and consider level 5 and below (not shown here), we find that all VARIETY-based splits separate the data in the same direction: IndE shows generally a greater preference for the prepositional variant than BrE, CanE, or SgE. This is especially true in those cases where VARIETY is interacting with NP expression type: if the recipient is non-pronominal (e.g. *John*), IndE favors the prepositional dative more than the other varieties. On the upper levels of the tree (level 4 and above), and in contrast to the other varieties, IndE also favors the prepositional dative when the recipient is longer than one word and the theme is up to four words long (Node 22). This ties in nicely with previous findings that IndE displays a higher proportion of prepositional datives than other varieties (e.g. Olavarría de Ersson and Shaw 2003; De Cuypere and Verbeke 2013; Mukherjee and Hoffmann 2006). Also consistent with our findings, De Cuypere and Verbeke (2013) have pointed out that IndE does not exhibit the tendency of aligning the constituents harmonically to the same degree as other varieties (short before long, pronoun before noun phrase, etc.) but "may have developed in a manner that differed greatly from the evolution of the other macro-regional varieties of English" (p. 180).

In sum, the length and expression type of the recipient/theme turn out to be the two most important factors in the upper branches of the tree, highlighting the high degree of interaction between pronominality and length. Also, when the theme is longer than one word, the interactions between the different factors reflect the influence of end-weight: when the recipient is shorter than the theme, the ditransitive variant is preferred (Nodes 10, 17). When both constituents have the same length (Node 21), the prepositional variant is preferred. Lastly, we find that when the theme is very short (one word), RECWORDLTH has very little influence.

To tease apart the effects of length and pronominality further, we computed another tree (not shown), this time ignoring all observations with pronominal constituents (N = 1,423). In this tree, VARIETY gains predictive power. Again, we find that the length of the theme is the single most reliable predictor, but among shorter themes (< four words) we find a significant effect of VARIETY. More than any other variety, IndE exhibits a greater overall preference for the prepositional dative variant, especially in cases where other varieties favor the ditransitive variant, e.g. in contexts involving proper noun themes and/or shorter recipients. In a nutshell, then, variety differences are more pronounced when both recipient and theme are not pronominal. The point is that, generally, processing constraints on the dative alternation tend to overshadow cross-varietal differences. When we control for the most powerful of these processing constraints, however, spoken IndE diverges from the other varieties in that it prefers prepositional datives even with long themes while speakers of BrE, SgE and CanE are more sensitive to theme length. This variety-based split can only be found in spoken language. In written texts, factors related to length measurements of theme and recipient are the single most important predictors. Whether this divergence between the varieties has to do with the level of formality merits future exploration.

CRF analysis performs slightly better than the conditional inference trees: C is 0.93 for both runs (compared to 0.86 for the tree), while predictive accuracy is 85.1% and 84.3%, thus slightly less than for the conditional inference tree, but still significantly better than the baseline of 68.2% ($p_{\text{binom}} < 0.001$). Figure 7 shows the relative ranking of model predictors obtained from the CRF analysis. The relative ranking in the CRF is mostly consistent with the single inference tree analysis, in that the top four predictors in the CRF ranking are represented in numerous high-level splits in Figure 6.

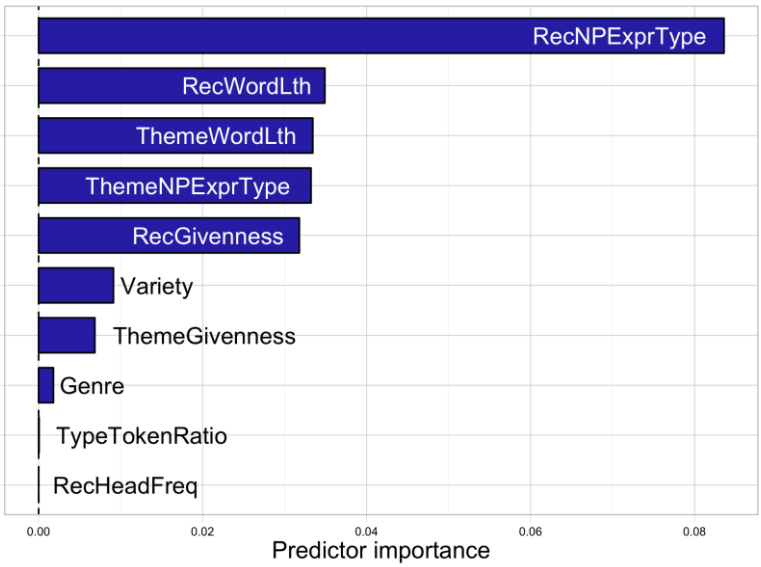

Figure 7: Predictor importance ranking for CRF analysis of the dative alternation

We then experimented with a second CRF analysis (not shown, *C*: 0.86, *N* = 1,423), this time ignoring all dative tokens with pronominal constituents. Cross-varietal differences are much more pronounced when both recipient and theme are realized as full noun phrases.

In conclusion, the length, pronominality, and givenness of both the recipient and theme influence the choice of dative variant across the board. Regional variability only becomes more prominent when pronominal NPs are ignored.

5. Discussion

In this section, we consider how the findings speak to the research questions laid out in Section 1.

Do the varieties of English we study here share a core probabilistic grammar? Yes, in the sense that there clearly are variety-independent, qualitative generalizations. For example, wherever we look in our data, longer constituents follow shorter constituents – a pattern that is known as the principle of end weight. We wish to add that in our view, weight effects are a part of grammar as well as symptomatic of processing demands – grammar and processing are not mutually exclusive. As for alternation-specific factors, we saw that in the particle placement alternation, directional PPs consistently favor the verb-object-particle order; in the genitive alternation, possessors ending in a final sibilant consistently favor the *of*-genitive; and in the dative alternation, pronominal themes consistently favor the prepositional dative variant. Hence the *effect directions* of these factors are stable across varieties of English. That being said, there seem to be interesting quantitative differences with regard to the *effect size* of the constraints on variation. For example, in the particle placement alternation it seems as though the effect of a directional PP following the target VP is weaker in IndE than in the other varieties we studied. By way of generalization, it seems that we find cross-variety differences of this kind only in those contexts where neither alternate is more or less difficult to process – for example, in the particle placement alternation we see a significantly lower proportion of the split variant in non-native varieties only when we restrict attention to very short

direct objects. We observe similar, though less robust, tendencies in the genitive and dative alternations (see sections 4.2.3 and 4.3.3).

Can ecology account for probabilistic similarity between varieties of English – for example, do we find a split between native and non-native varieties of English? In some cases, the four varieties we study in this paper do indeed tend to pattern along native versus non-native (or Inner Circle versus Outer Circle or ENL versus ESL) lines. Consider that in both the particle placement and, to a lesser extent, the genitive alternation, the conditional inference trees pit BrE and CanE against IndE and SgE. In the dative alternation, by contrast, IndE is set apart from a cluster joining the other varieties. Because our results are not entirely conclusive, further research is needed.

Do the alternations investigated differ in terms of their probabilistic sensitivity to variety effects? The existence of a core grammar (see above) notwithstanding, the three alternations we study differ as to how amenable they are to "probabilistic indigenization", as it were. We define probabilistic indigenization as the process whereby stochastic patterns of internal linguistic variation are reshaped by shifting usage frequencies in speakers of post-colonial varieties. To the extent that patterns of variation in a new variety A, e.g. the probability of item x in context y, can be shown to differ from those of the mother variety, we can say that the new pattern represents a novel, if gradient, development in the grammar of A. These patterns need not be consistent or stable (especially in the early stages of nativization), but they nonetheless reflect the emergence of a unique, region-specific grammar. Of the three phenomena we investigate, the particle placement alternation exhibits the most robust variety effects – in fact, variety is ranked as the single most important predictor of particle placement choice by CRF analysis. The conditional inference trees suggest that variety effects seem least pronounced in the genitive alternation (at least on the basis of the constraints we currently include in multivariate modeling). Building on Schneider's observation that lexico-grammar is a prime target of early-stage indigenization (Schneider 2003, 249), we tentatively offer the following generalization: the more tightly associated a given syntactic alternation is with concrete instantiations involving specific lexical items – consider verb slots in the particle placement and dative alternation – the more likely it is to exhibit cross-varietal indigenization effects (Hoffmann 2014; Grafmiller and Röthlisberger forthcoming).

Beyond these issues, we note that according to CRF analysis, in all datasets we study, variety effects appear to be more important than genre differences.

6. Conclusion and next steps

We fully concur with Adger and Trousdale that variation is, or should be, a "core explanandum" (2007, 274) in linguistic theorizing. As a consequence, the project on which we have embarked in Leuven is a large-scale, comparative endeavor to test the scope and limits of probabilistic variation in World Englishes. The innovative potential of this work derives from its emphasis on the probabilistic, usage- and experience-based nature of linguistic variation: we assume that language users implicitly learn the probabilistic effects of constraints on variation by constantly (re-)assessing input of spoken and written discourses they engage in throughout their lifetimes – and crucially, this input is likely to differ across different speech communities and varieties of English. We thus combine an interest in probabilistic modeling of variation across World Englishes with an interest in socially and cognitively

contextualized language usage. Thanks to its multi-disciplinary and cross-pollinating orientation, the work sketched here is hoped to bridge gaps between different strands of theoretically oriented usage-based linguistics, thus facilitating fertile interface explorations.

Of course, the analyses reported here are just a first step. Work is underway in Leuven to analyze data from no fewer than nine varieties of English from around the world: in addition to BrE, CanE, IndE, and SgE – which we analyze in the present study – we will also be including Irish English, New Zealand English, Hong Kong English, Jamaican English, and Philippine English. In addition to the International Corpus of English (ICE) we are working on tapping into the Corpus of Global Web-based English (GLoWbE; Davies and Fuchs 2015), with the goal of adding web-based language to the array of text types sampled in ICE. As far as the probabilistic constraints subject to analysis are concerned, the case studies reported in this paper have relied on factors that are fairly easy to annotate, but we are currently engaging in meticulous hand-coding to annotate syntactic choices for factors such as information status and – in particular – NP animacy. And as for the type of evidence we consider, corpus analysis takes center stage in our project (as it did in the present paper), but we plan to spot-check the cognitive robustness of the corpus-derived probabilities via rating experiments along the lines of Bresnan (2007) and Bresnan and Ford (2010), who showed that language users' acceptability ("naturalness") intuitions about syntactic choices match probabilities as calculated in a corpus-based regression model.

Note

The first-named author is the principal investigator of the project. Jason Grafmiller is the investigator of the particle placement study, Benedikt Heller is the investigator of the genitive study, and Melanie Röthlisberger is the investigator of the dative study.

References

Adger, David, and Graeme Trousdale. 2007. “Variation in English Syntax: Theoretical Implications.” *English Language and Linguistics* 11 (02): 261. doi:10.1017/S1360674307002250.

Baayen, R. Harald, Anna Endresen, Laura A. Janda, Anastasia Makarova, and Tore Nesset. 2013. “Making Choices in Russian: Pros and Cons of Statistical Methods for Rival Forms.” *Russian Linguistics* 37 (3): 253–91. doi:10.1007/s11185-013-9118-6.

Behagel, Otto. 1909. “Beziehungen Zwischen Umfang Und Reihenfolge von Satzgliedern.” *Indogermanische Forschungen* 25: 110–42.

Bernaisch, Tobias, Stefan Th. Gries, and Joybrato Mukherjee. 2014. “The Dative Alternation in South Asian English(es): Modelling Predictors and Predicting Prototypes.” *English World-Wide* 35 (1): 7–31. doi:10.1075/eww.35.1.02ber.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Bod, Rens, Jennifer Hay, and Stefanie Jannedy, eds. 2003. *Probabilistic Linguistics*. Cambridge, MA: MIT Press.

Bresnan, Joan. 2007. “Is Syntactic Knowledge Probabilistic? Experiments with the English Dative Alternation.” In *Roots: Linguistics in Search of Its Evidential Base*, edited by Sam Featherston and Wolfgang Sternefeld, 75–96. Berlin: Mouton de Gruyter.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and Baayen Harald. 2007. “Predicting the Dative Alternation.” In *Cognitive Foundations of Interpretation*, edited by G Boume, I Kraemer, and J Zwarts, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, Joan, and Marilyn Ford. 2010. “Predicting Syntax: Processing Dative Constructions in American and Australian Varieties of English.” *Language* 86 (1): 168–213. doi:10.1353/lan.0.0189.

Bybee, Joan, and Paul Hopper. 2001. *Frequency and the Emergence of Linguistic Structure*. Amsterdam: Benjamins.

Cappelle, Bert. 2005. “Particle Patterns in English: A Comprehensive Coverage.” Ph.D. Thesis, Leuven, Belgium: K.U. Leuven.

Davies, Mark, and Robert Fuchs. 2015. “Expanding Horizons in the Study of World Englishes with the 1.9 Billion Word Global Web-Based English Corpus (GloWbE).” *English World-Wide* 36 (1): 1–28. doi:10.1075/eww.36.1.01dav.

De Cuypere, Ludovic, and Saartje Verbeke. 2013. “Dative Alternation in Indian English: A Corpus-Based Analysis.” *World Englishes* 32 (2): 169–84. doi:10.1111/weng.12017.

Ehret, Katharina, Christoph Wolk, and Benedikt Szmrecsanyi. 2014. “Quirky Quadratures: On Rhythm and Weight as Constraints on Genitive Variation in an Unconventional Data Set.” *English Language and Linguistics* 18 (02): 263–303. doi:10.1017/S1360674314000033.

Gahl, Susanne, and Susan Garnsey. 2004. “Knowledge of Grammar, Knowledge of Usage: Syntactic Probabilities Affect Pronunciation Variation.” *Language* 80: 748–75.

Gahl, Susanne, and Alan C.L. Yu. 2006. *Special Theme Issue: Exemplar-Based Models in Linguistics*. The Linguistic Review. Mouton de Gruyter.

Garretson, Gregory, M. Catherine O’Connor, Barbora Skarabela, and Marjorie Hogan. 2004. “Coding Practices Used in the Project Optimality Typology of Determiner Phrases.” http://npcorpus.edu/documentation/index.html.

Grafmiller, Jason. 2014. “Variation in English Genitives across Modality and Genres.” *English Language and Linguistics* 18 (03): 471–96. doi:10.1017/S1360674314000136.

Grafmiller, Jason, and Melanie Röthlisberger. forthcoming. “Construction Grammar Goes Global: Syntactic Alternations, Schematization, and Collostructional Diversity in World English(es)”

Gries, Stefan Th. 2003. *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*. New York: Continuum Press.

Harrell, Frank E. 2001. *Regression Modeling Strategies With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York, NY: Springer New York. http://dx.doi.org/10.1007/978-1-4757-3462-1.

Harrell, Frank E. Jr. 2014. *Hmisc: Harrell Miscellaneous. R Package Version 3.14-6* (version 3.14-6). http://CRAN.R-project.org/package=Hmisc.

Hinrichs, Lars, and Benedikt Szmrecsanyi. 2007. “Recent Changes in the Function and Frequency of Standard English Genitive Constructions: A Multivariate Analysis of Tagged Corpora.” *English Language and Linguistics* 11: 437–74. doi:10.1017/S1360674307002341.

Hoffmann, Thomas. 2014. “The Cognitive Evolution of Englishes: The Role of Constructions in the Dynamic Model.” In *Varieties of English Around the World*, edited by Sarah Buschfeld, Thomas Hoffmann, Magnus Huber, and Alexander Kautzsch, 160–80. Amsterdam: John Benjamins Publishing Company. https://benjamins.com/catalog/veaw.g49.10hof.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. 2006. “Unbiased Recursive Partitioning: A Conditional Inference Framework.” *Journal of Computational and Graphical Statistics* 15 (3): 651–74. doi:10.1198/106186006X133933.

Kachru, Braj B., ed. 1992. *The Other Tongue: English across Cultures*. 2nd ed. English in the Global Context. Urbana: University of Illinois Press.

Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Philadelphia Press.

———. 1982. “Building on Empirical Foundations.” In *Perspectives on Historical Linguistics*, edited by Winfred Lehmann and Yakov Malkiel, 17–92. Amsterdam, Philadelphia: Benjamins.

Levy, Roger. 2008. “Expectation-Based Syntactic Comprehension.” *Cognition* 106: 1126–77. doi:10.1016/j.cognition.2007.05.006.

MacDonald, Maryellen C. 2013. “How Language Production Shapes Language Form and Comprehension.” *Frontiers in Psychology* 4: 1–16. doi:10.3389/fpsyg.2013.00226.

Mesthrie, Rajend, and Rakesh Mohan Bhatt. 2008. *World Englishes: The Study of New Linguistic Varieties*. Key Topics in Sociolinguistics. Cambridge, UK ; New York: Cambridge University Press.

Mukherjee, Joybrato. 2007. “Steady States in the Evolution of New Englishes: Present-Day Indian English as an Equilibrium.” *Journal of English Linguistics* 35 (2): 157–87.

Mukherjee, Joybrato, and Sebastian Hoffmann. 2006. “Describing Verb-Complementational Profiles of New Englishes: A Pilot Study of Indian English.” *English World-Wide* 27: 147–73. doi:10.1075/eww.27.2.03muk.

Olavarría de Ersson, Eugenia, and Philip Shaw. 2003. “Verb Complementation Patterns in Indian Standard English.” *English World-Wide* 24: 137–61.

Osselton, Noel. 1988. “Thematic Genitives.” In *An Historic Tongue: Studies in English Linguistics in Memory of Barbara Strang*, edited by Graham Nixon and John Honey, 138–44. London: Routledge.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London and New York: Longman.

Rohdenburg, Günter. 1996. “Cognitive Complexity and Increased Grammatical Explicitness in English.” *Cognitive Linguistics* 7 (2): 149–82. doi:10.1515/cogl.1996.7.2.149.

Rosenbach, Anette. 2014. “English Genitive Variation – the State of the Art.” *English Language and Linguistics* 18 (02): 215–62. doi:10.1017/S1360674314000021.

Schilk, Marco, Tobias Bernaisch, and Joybrato Mukherjee. 2012. “Mapping Unity and Diversity in South Asian English Lexicogrammar: Verb-Complementational Preferences across Varieties.” In *Mapping Unity and Diversity World-Wide: Corpus-Based Studies of New Englishes*, edited by Marianne Hundt and Ulrike Gut, 137–65. Amsterdam: John Benjamins.

Schneider, Edgar. 2007. *Postcolonial English: Varieties Around the World*. Cambridge University Press.

Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. 2008. “Conditional Variable Importance for Random Forests.” *BMC Bioinformatics* 9 (1): 307. doi:10.1186/1471-2105-9-307.

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. “Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution.” *BMC Bioinformatics* 8 (1): 25. doi:10.1186/1471-2105-8-25.

Tagliamonte, Sali. 2001. “Comparative Sociolinguistics.” In *Handbook of Language Variation and Change*, edited by Jack Chambers, Peter Trudgill, and Natalie Schilling-Estes, 729–63. Malden and Oxford: Blackwell.

Tagliamonte, Sali, and Harald Baayen. 2012. “Models, Forests and Trees of York English: ‘Was/were’ Variation as a Case Study for Statistical Practice.” *Language Variation and Change* 24: 135–78.

Wasow, Thomas, and Jennifer Arnold. 2003. “Post-Verbal Constituent Ordering in English.” In *Determinants of Grammatcal Variation in English*, edited by G. Rohdenburg and B. Mondorf, 119–54. Amsterdam: de Gruyter.

Wolk, Christoph, Joan Bresnan, Anette Rosenbach, and Benedikt Szmrecsanyi. 2013. “Dative and Genitive Variability in Late Modern English: Exploring Cross-Constructional Variation and Change.” *Diachronica* 30 (3): 382–419. doi:10.1075/dia.30.3.04wol.

Zwicky, Arnold M. 1987. “Suppressing the Zs.” *Journal of Linguistics* 23: 133–48.